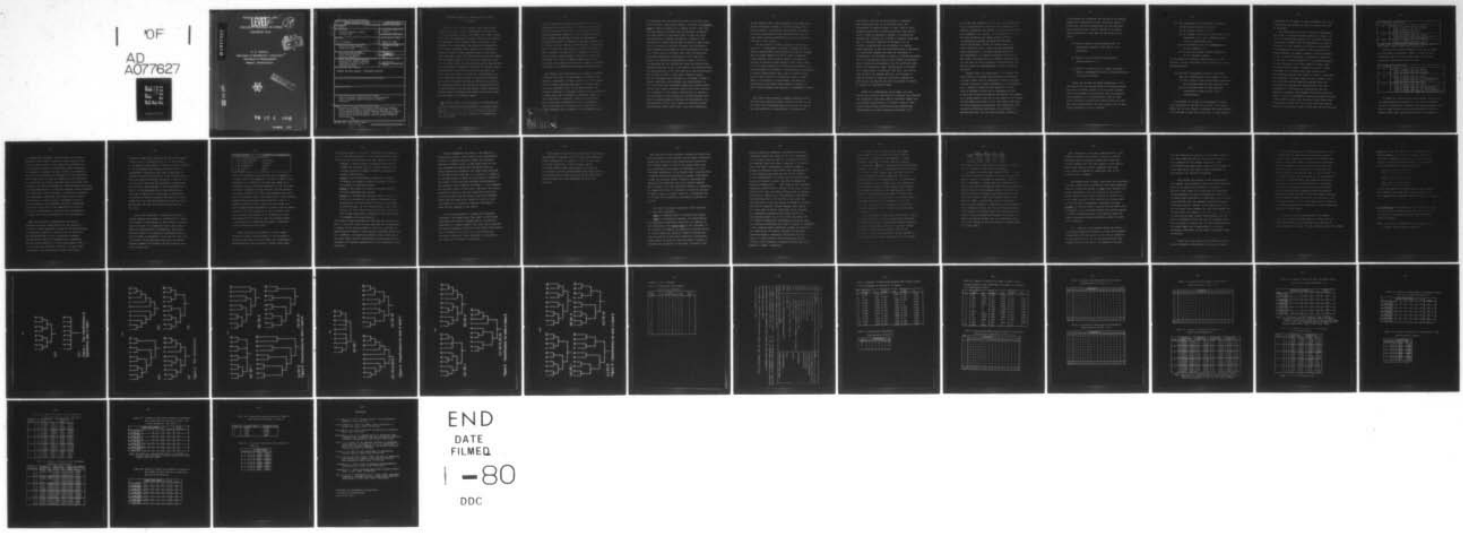
AD-A077 627 MASSACHUSETTS UNIV AMHERST DEPT OF MATHEMATICS AND S--ETC F/G 12/1
SIMILARITY MEASURES ON BINARY ATTRIBUTE DATA.(U)
NOV 79 M F JANOWITZ N00014-79-C-0629
UNCLASSIFIED TR-J7901 NL



END
DATE FILMED
-80
DDC

Technical Report Number J7901

LEVEL II

12 B.S.

SIMILARITY MEASURES ON BINARY ATTRIBUTE DATA

M. F. Janowitz
Department of Mathematics & Statistics
UNIVERSITY OF MASSACHUSETTS
Amherst, Massachusetts

NOVEMBER 1979

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: TR-J7901	2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE (and Subtitle): Similarity Measures on Binary Attribute Data.	5. TYPE OF REPORT & PERIOD COVERED: Technical rept.
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): M. F. Janowitz	8. CONTRACT OR GRANT NUMBER(s): N00014-79-C-0629
9. PERFORMING ORGANIZATION NAME AND ADDRESS: University of Massachusetts, Amherst, MA 01003	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: 721405
11. CONTROLLING OFFICE NAME AND ADDRESS: Procuring Contracting Officer, Office of Naval Research, Arlington, VA 22217	12. REPORT DATE: November, 1979
	13. NUMBER OF PAGES: 42
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office): Office of Naval Research Resident Representative, Harvard University, Gordon McKay Laboratory, Room 113, Cambridge, MA 02138	15. SECURITY CLASS. (of this report): Unclassified
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

APPROVED FOR PUBLIC RELEASE: DISTRIBUTION UNLIMITED.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Numerical taxonomy, Cluster analysis, Similarity Measure, Special clustering, Optimality measures, Cophenetic correlation

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

The ability of commonly used similarity coefficients to recapture natural classifications in the presence of various types of errors is investigated by means of several computer simulations of a certain model of the classification problem. The goal is to establish the relative merits of the phylogenetic versus the phenetic methods of classification.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

S/N 0102-LF-014-6601

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

411 377

LB

Similarity Measures on Binary Attribute Data

M. F. Janowitz *

Farris (1977) compares phenetic with phylogenetic taxonomy and claims to have established the superiority of the phylogenetic system. Upon reading his paper I found that I did not agree with a certain portion of his reasoning (Janowitz, 1979), though I took no position on the validity of his conclusion. The inevitable vituperative reply occurs in Farris (1979). Though it is tempting to respond in detail to the assertions made in this reply, sober reflection leads me to the conclusion that the reader will be better served if I do not do so. After all, this is not a debate where the person who makes the most skillful use of the English language is declared the winner; rather, there are some facts about cluster analysis that both Farris and I are trying to establish. Unfortunately, our basic underlying principles are so drastically different that we seem destined to arrive at differing conclusions. The reader must decide for himself the relative merits of our two positions, and for that reason I shall devote the present paper to clarifying and expanding upon the points that were made in my original paper.

§1. Farris (1979) is quite correct in mentioning some mistakes in arithmetic that occurred in my earlier paper

---

*Research supported by ONR Contract N00014-79-C-0629 as well as by grants from the University of Massachusetts Computer Center.

(some of these were typographical errors, and some pure carelessnes on my part), and then proceeds to demonstrate precisely the point that I was trying to make. By considering artificial examples one can produce situations where phemetic methods are superior to phylogenetic methods and conversely. Indeed, he goes even further than this when he states (Farris, 1979:206) that "By suitable choice of data sets, any method of classification can be made to appear as unstable as one pleases." It follows that one cannot draw a general conclusion from the consideration of a specific example. Since this is one of the points I was trying to make with my example (Janowitz, 1979:197), it is gratifying that Farris agrees with me.

§ 2. What is the significance of the example provided by Farris (1977:837)? For the reader's convenience the example is reproduced in Table 1 and Fig. 1. This example involves binary character data where 1 is always taken to be the preferred state. If, as Farris claims, his special clustering is in general superior to raw clustering, I would expect this superiority to manifest itself in this very specific situation. Since it is especially easy to see what is happening in such a situation, I cannot lose anything by restricting my attention to binary character data in which 1 is the preferred state -

an assumption that was implicitly made in my earlier paper also. In Farris' terminology (Farris, 1979:202) the reference point is taken to possess all 0 states; he and I then seem to agree (Farris, 1979:202) that his measure s represents the simple matching coefficient (DC1 in Table 2), while his coefficient of special similarity a reduces to the coefficient of Russel and Rao (DC4 in Table 2). When either s or a is applied to the data of Table 1 and followed by UPGMA clustering, the result is the classification of Fig. 1(a). Farris argues (1977:832) that since the associated cophenetic correlation coefficient is 1.00, and since "each of the features corresponding to state 1 of one of the variables has its distribution among the terminal taxa perfectly described by one of the taxa of the classification", the classification of Fig. 1(a) must be a good classification for the data of Table 1. With this I heartily concur. When the characters are replicated as indicated in Table 1, special similarity still produces the classification of Fig. 1(a), while raw clustering produces the classification of Fig. 1(b). Farris (1977:834) argues that since "replication introduces no new types of distribution of features into the data, but only alters the relative frequency with whice the various types of features are represented", the classification of Fig. 1(a) "remains a perfectly natural classification after replication, since the distribution

of the values of any of the variables of the data may be exactly described in terms of one of the taxa of the classification." He concludes that phenetic similarity clustering can produce a poor classification, even in cases where a natural classification can be recognized. With this I also concur.

But the example of Table 1 involves a fully congruent set of characters, so there is already at hand a natural classification, and one does not need to use any type of cluster method to find it. Does it therefore matter whether a given cluster method is superior or inferior to another cluster method on this type of data? Is it not more pertinent to investigate the behaviour of cluster methods with respect to their ability to reconstruct a natural classification from a set of characters that are not fully congruent? In sections 3 and 4 I shall examine this question in some detail, and present a statistically significant large number of examples in which phenetic methods are superior to phylogenetic methods.

§3. What is the situation regarding characters that are not fully congruent? Farris (1979:206) states in part that "Even for incongruent data, it goes much too far to say that sensitivity to replication is irrelevant to

naturalness, although the relationship is certainly more complicated than in the congruent case. The relationship is complex enough, in fact, that it would probably be difficult to devise artificial data through which it might be adequately studied." But artificial examples are important for a number of reasons. Among them are: (i) Artificial data can be designed so as to isolate a property that one is studying in a simple enough situation so that one can analyze what is happening. (ii) Artificial examples can serve as indicators of possibilities for the truth or falsity of various hypotheses. (iii) If a proposition about cluster methods is generally true, it must also be generally true for artificial data. Hence if a proposition seems statistically to be false for a large number of artificial data sets, one must at least question its general validity. At any rate, no harm can be done by examining artificial examples, as long as one uses them in the spirit of gaining intuition as opposed to establishing facts.

Before any investigation can be begun, one must decide upon a precise definition of a natural classification for incongruent input data. Farris (1977:829) states "the most natural classification in Gilmour's sense is that classification whose constituent groups describe the distributions of as many features as posiible." If I

am to take this statement at face value, it appears that Farris is advocating the selection of a maximal set of congruent characters, and deeming the classification that they represent to be natural. It is interesting to note that this represents the rather elegant method of "compatability analysis" advocated by Estabrook, Johnson and McMorris (1976). But Farris goes on to state (1977:830) that "...while the correspondence between membership of a Gilmour-natural taxon and the distribution of a feature considered described by that taxon is allowed not to be perfect, the taxon cannot very well be said to describe the distribution of a feature unless the correspondence is kept as close as possible."

Though I plan in a later paper to investigate the above notion of a natural classification, for present purposes I shall find it convenient to adopt a slightly different viewpoint. Before proceeding, it is imperative that I carefully establish that framework in which I shall be working. Without such a framework, any discussion can be both confusing and misleading. For example, it is quite possible that one of the causes of the dispute between Farris and myself is the fact that we have made different underlying assumptions, and consequently have arrived at differing conclusions. I am working in a mathematical model for the classification problem, so

I must ignore any information not implied by my abstract mathematical assumptions; on the other hand, Farris is making some biological assumptions, and can therefore reach conclusions not directly implied by his mathematical assumptions. Here then are the axioms for my model:

A1. There is given a finite nonempty set P of operational taxonomic units (OTU's) to be classified.

A2. There is a unique desired hierarchical classification of P.

A3. There is a finite set A of binary (presence-absence) characters from which the classification of A2 may be deduced.

Unless one can make the above assumptions, it is difficult to see how the application of cluster analysis can lead to a useful result. It is of course possible that if one changes the criteria by which one measures the desirability of a classification, then A2 might allow a different hierarchical classification, but with each such classification, one must assume a set of characters from which it may be deduced.

A4. The investigator now enters the picture and introduces some errors as follows:

(i) He chooses a subset A' of A.

(ii) He misreads a portion p of the states of the characters in A', due to sampling errors, errors in measurement, etc.

(iii) He introduces a set B of characters that are not members of A.

(iv) He may even miscode a small proportion q of his character states.

If the concept of a natural classification is to be included, it seems reasonable that it might do so in the following manner:

A5. The set of characters specified in A3 shall make the classification of A2 natural in that:

(i) Each cluster of the classification represents the preferred state of some character, and

(ii) The preferred state of each character corresponds to some cluster in the classification.

The problem now facing the investigator is clear. Given the errors he has made in A4, what is the cluster method that is most likely to reproduce the classification of A2? One way to test this is by means of some computer

simulations of the types of errors described in A4. Such a simulation was carried out, and its results will now be described.

The first thing to notice is that the investigator has no way of knowing the exact contents of either A or A'; all he has is his version of A' (possibly including some errors) together with the characters in B. The behaviour of cluster methods with respect to replication of characters and with respect to the introduction of error characters is clearly relevent to what transpires in A4. A cluster method that is relatively insensitive to the type of error descibed in A4 is clearly desirable in this situation. As a first attempt to obtain some intuition for what might be happening, I took 7 random characters on a 3 element set, rank ordered the 12 dissimilarity measures that appear in Table 2, replicated the characters in various ways, and then checked to see how often the rankings remained unchanged after the dissimilarity measures were recalculated. Each trial consisted of 5 different such replications, and the portion of successes was recorded. The results appear in Table 3. Using a nonparametric signs test (Gibbons:1976,94) I then tried to assess the statistical significance that DCi performed better than DCj in this particular simulation. At a significance level of 90%, here were the results:

No. Trials	T	Conclusion
25	10	DC10 is better than all others DC7 is better than DC5 and DC9 DC3 is better than DC5 DC11 is better than DC9
25	5	DC10 is better than all others DC3,DC4,DC7,DC8,DC11 are better than DC1 DC6 is better than DC1,DC5,DC12
25	2	DC10 is better than all but DC5 DC1 is worse than all others

In each trial a character is replicated a random number of times with a number chosen from 0,1,2,...,T.

Similar results were performed on the data in Table 4 with the results appearing in Table 5. Using the same statistical test as for the random data here are the results at a significance level of 90%:

No. Trials	T	Conclusion
25	10	DC9 is better than all others DC4 is better than all but DC9 DC1 is better than DC6,DC7,DC11
25	5	DC9 is better than all others DC4 is better than all but DC1,DC2,DC3,DC9 DC1 is better than all but DC3,DC4,DC9 DC2,DC3 are better than DC10
25	2	DC4 is better than all others DC1 is better than all but DC3,DC4,DC8,DC9 DC3 is better than all but DC1,DC4,DC9 DC9 is better than all but DC1,DC3,DC4,DC8 DC8 is better than than DC2, DC5,DC6,DC10,DC12

A number of conclusions may now be drawn. First of all, the results seem data dependent, so one must be extremely cautious in drawing any conclusion from them. Secondly, though there is some evidence that special similarity (DC4) may perform better than simple matching (DC1), the evidence is

not overwhelming. Thirdly, one should bear in mind that those DC's that tended to produce a large number of ties would not show up well in this particular simulation, as ties are very unlikely to remain stable under replication of characters. Fourthly, the data in Table 4 is extremely sensitive to replications in characters, so the conclusions that I drew from this data are probably not as significant as the earlier conclusions that were based upon random data. Finally, in all but one simulation, a dissimilarity measure was best that happened also to treat the character states symmetrically. This would tend to indicate a superiority of phenetic over phylogenetic techniques. The results are of course inconclusive, but they do make it reasonable for me to focus my attention on DC1,DC4,DC9 and DC10; furthermore, they do cast some doubt on the superiority of DC4 as a measure of dissimilarity. A more realistic simulation is in order, and it will be described in the next section.

§4. How well does a dissimilarity coefficient recapture a natural classification in the presence of the type of error described in Axiom A4? Before attempting to answer this question, a few words are in order concerning the ability of a DC to recapture a natural classification from fully congruent input characters. The test classifications I have chosen appear in Fig. 2, and are taken from D'Andrade (1978:65). In view of the discussion of the last section, I shall restrict my

attention to DC1, DC4, DC9 and DC10. The cluster method I shall use is u-clustering with u = .5. This method is described in some detail in Janowitz (1979a). Basically, it is an agglomerative hierarchical method that is implemented by merging at each level those pairs of clusters for which at least half of the possible links have been made. Despite the discussion of section 2, it is not at all clear that DC4 necessarily produces the most faithful representation of the classifications in Fig. 2. If one represents these classifications as indicated in Tables 6-9, for example, an examination of Figures 3-6 seems to indicate that DC10 best represented Fig. 2(a), DC4 and DC9 were best for Fig. 2(b), DC4 was best for Fig. 2(c), while DC1 was best for Fig. 2(d).

Since I am attempting to isolate the behavior of cluster methods with respect to their ability to recapture natural classifications in the presence of certain types of errors, it is essential that I start with a character representation of hierarchical trees that allows the trees to be perfectly recaptured by all of the DC's under consideration. To see how this goes, and incidentally to see what went wrong with the earlier attempt, consider the clusters that are found at each level of Fig. 2(a):

Level	Clusters	Character Representation
0	1,2,3,4,5,6,7,8,9,10	10-19
1	12,34,56,78,9-10	1,2,3,4,5
2	1-4,56,78,9-10	6,3,4,5
3	1-6,78,9-10	7,4,5
4	1-8,9-10	8,5
5	1-10	9

Notice that Character 3 must be used twice, Character 4 is used 3 times, and Character 5 4 times. In Table 6 this information is contained in the last row, where it is indicated how many times each character must be replicated in order to perfectly reflect the data. Similar observations apply to Tables 7-9. When the characters are replicated as indicated, each of DC1 through DC12 will perfectly represent the appropriate classifications of Fig. 2. A moments reflection should convice you that though our earlier scheme seemed reasonable, it had no hope of success because it did not reflect all of the clusters at each level at which they appear. It is interesting to note that Farris(1979) did not run into this problem because no replications of characters are needed to perfectly represent the tree of Fig. 1(a).

Having disposed of the manner in which I shall represent a natural classification, I can now begin to introduce some errors. For each data set I introduced 3 extra characters that consisted of random binary data.

I then applied DCi (i = 1,4,9,10) followed by u-clustering (u = .5). The goodness of fit of the resulting classification to the original classification was then measured in 4 ways.

Method 1. A product moment correlation was computed between the output DC d of the cluster method and the original unperturbed DC $\hat{d}'$ that reflected perfectly the desired classification.

Method 2. Each classification represents a certain collection of desired clusters. The number of missing but desired clusters was counted.

Method 3. The number of extraneous clusters added by the error characters was counted.

Method 4. A distance was computed between d and d' of Method 1 by counting up the number of unordered pairs $\{\{a,b\},\{x,y\}\}$ for which either (i) $d(a,b) = d(x,y)$ and $d'(a,b) \neq d'(x,y)$, or (ii) $d(a,b) \neq d(x,y)$ and $d'(a,b) = d'(x,y)$, or (iii) $d(a,b) < d(x,y)$ and $d'(a,b) > d'(x,y)$. This is then normalized by dividing by the total number of unordered pairs $\{\{a,b\},\{x,y\}\}$.

There were 10 trials performed on each data set and with each DCi (i = 1,4,9,10) though in some cases DC9 was not considered. A summary of the results appears in Table 10, with Table 11 summarizing the number of times that DCi performed better than DCj in Method 1. Note that the evidence points to DC4 being the worst choice and DC10 the best choice with respect to the recapture of a natural classification in the presence of error characters.

Having investigated the effect of the addition of random error vectors, I then turned to the consideration of errors in reading character states. The data sets contained in Tables 6,7,8,9,1 were again used, but now each character was doubled and a 5% random error in reading states was introduced. Ten trials were performed on each data set and each DCi (i = 1,4,8,10). A summary of the results appears in Table 12, with Table 13 indicating the number of times that DCi performed better than DCj with respect to the criterion of Method 1. I also compared DC4 with DC10 using 5 trials with a 15% error, the results appearing in Table 14. The only significant conclusion that can be drawn from this data is that DC10 performed better than the other DC's, though it should be noted that there was one instance (with data from Table 6) where DC4 was superior to the others.

As a final simulation, I doubled each character, introduced a 5% random error, then added 6 random error characters, and finally discarded 10% of the resulting characters in a random fashion. The intent was to simulate the errors described in Axiom A4 of the model. With respect to the criteria of Methods 1 and 4, the results are tabulated in Table 15. Notice that DC4 is sonsistently the worst choice of those I considered.

With respect to all of these simulations, the evidence overwhelmingly indicates that a DC that treats character states symmetrically is superior to Farris' coefficient of special similarity. In short, though special similarity performs well with respect to its ability to recapture a natural classification from fully congruent input characters, it does a poor job in the more general situation where the input characters are not fully congruent.

§ 5. The role of the cophenetic correlation coefficient. This coefficient is the ordinary product moment correlation between the input and output dissimilarity coefficients of a cluster method. In both of his papers Farris uses this coefficient as a measure of performance of a dissimilarity coefficient. In my earlier paper (Janowitz,1979) I objected to this and presented an extreme illustration of what can go wrong. Farris (1979:207) questioned the validity of my objection, and quite clearly explained why he wishes to use this particular measure of optimality. The reader will find it instructive to read this section of Farris' paper. In view of this, it is appropriate for me to begin this discussion by restating my earlier objection.

The type of clustering algorithm under discussion is a two stage algorithm:

Stage 1. Character data $\longrightarrow$ dissimilarity measure

Stage 2. Dissimilarity measure $\longrightarrow$ classification.

The cophenetic correlation coefficient is a measure of the optimality of the second stage of the algorithm. It can be used, for example, to compare the relative merits of single linkage and complete linkage clustering when they are applied to the same dissimilarity coefficient. Farris makes the error of using this Stage 2 measure to determine the optimality of the Stage 1 portion of the

process. Since the cophenetic correlation coefficient completely ignores the nature of the input characters, it can hardly decide how well the intermediate DC fits the original input data. What I suggested in my earlier paper, and what I now suggest, is that one wants a measure of optimality that relates the intermediate DC to the original input data, and not to the ultimate output classification. Of course this is precisely what was done in Section 4, and I now wish to compare the results in Tables 10-15 with those of the cophenetic correlation coefficient. A summary of the results occurs Tables 16-18. They are quite different from the results that occur in Tables 10-15. The cophenetic correlation indicates that DC9 is the best DC to use, while the earlier results (based upon actual performance) indicate that DC10 is superior. Notice, however, that despite the claims made by Farris (1977,1979) to the contrary, the cophenetic correlation coefficient for DC1 is consistently higher than it is for DC4 (See Table 16). To further illustrate the discrepencies caused by using the cophentic correlation as a Stage 1 measure of optimality, I ran a product moment correlation between the values of this measure and the Method 1 measure that directly determines Stage 1 optimality. The results occur in Table 19. They provide a vivid illustration of the collapse of the cophenetic correlation coefficient as a measure of Stage 1 optimality.

As a final item in this section, the reader should recall that in both of his papers, Farris stresses the importance of real examples. Indeed, in Farris (1979:212) he states, "No-one, however, has produced any analyses of real cases that give results different from mine. ... If pheneticists wish to dispute my findings, I would submit, then they can only do so through evidence from real organisms". Now I am a mathematician and not a taxonomist, so it would hardly be appropriate for me to attempt any such analysis of real organisms. However, I did compare raw clustering with special similarity on two real examples. Using data from Watson, Williams and Lance (1966:495) and using + as denoting presence with - as absence of each character, and taking as a reference point an OTU whith each state as -, the following values of the cophenetic correlation coefficient were obtained via $u = .5$ clustering:

DC1 = .8996, DC4 = .7925, DC9 = .9621, DC10 = .9162.

The second data set I chose was from Ferris and Whitt (1978:195). This time I used single-linkage clustering, $u = .5$ clustering, and a version of complete linkage clustering. I took as my reference point an OTU possessing none of the characters (as was intended by the authors of the source material). The results were:

	DC1	DC4	DC9	DC10
single	.8412	.7883	.8563	.6325
u = .5	.8701	.7289	.8684	.734
complete	.8254	.6144	.6948	.7303

Note that in both of these examples, special similarity does not perform as well as raw clustering!

In order to see why these results differ so much from those reported by Farris, one might consider the manner in which he assigned his reference points. In Farris (1977:836) he states that "the reference point used in each special similarity analysis being arbitrarily selected as one of the terminal taxa of the data"; in Farris (1979:210) he reports that the reference point was "taken simply as the first terminal taxon of the data set." When I applied this choice of reference point to the 2 examples, I obtained results comparable to those claimed by Farris. But this artificially asserts which state of a character shall be informative with no regard to the biological significance of the data. The reader must decide whether this is a reasonable thing to do. This situation will be explored in greater depth in a later paper.

§6. Conclusion. In my paper (Janowitz,1979) I very carefully refrained from taking a position on Farris' contention that phylogenetic methods are better than phenetic methods. All I did was attempt to show that in view of some flaws in Farris' reasoning, the issue should still be regarded as unresolved. This is the position I still maintain.

For certain types of data I have shown the superiority of Yule's coefficient (a coefficient that treats character states in a symmetric manner) to <u>both</u> simple matching and special similarity. I did not show and do not wish to assert that this makes Yule's coefficient better to use on all data. However, if the data has the property that for each OTU there corresponds a character possessed <u>uniquely</u> by that OTU, then my results indicate that Yule's coefficient is at least worth trying. The relative merits of special similarity versus simple matching were left largely unresolved, though the evidence tended to favor simple matching.

In addition to an argument based upon logic, I gave detailed empirical evidence as to why the cophenetic correlation coefficient should not be used to determine the relative merits of one dissimilarity measure over another. This was done within the framework of a model

that was introduced in Section 3 of the paper. There is much work remaining, and the results seem to raise as many questions as they asnwer. Hopefully, I have demonstrated that despite many claims to the contrary, no clear case has yet been made for the superiority of phylogenetic over phenetic clustering.

I also raised the question of the representation of a natural classification by means of binary character data. If one adopts the convention I introduced in Section 3, then there is no difficulty in arriving at a representation, but such a representation would not be stable under replication of characters; even for fully congruent data sets, the arguments presented by Farris (1977:833) would be invalid. On the other hand, if one does not adopt such a scheme, one need only consider the classifications of Figure 6 to see what can go wrong. The clusters are the same in each classification, but they arise at differing levels. How does one represent this? Must one agree to identify all 4 of the classifications? The reader might wish to ask whether it is either necessary or desirable to even search for natural classifications!

I spent some time examining the question of why my results differed so drastically from those of Farris.

As a possible explanation, I presented examples in which I obtained results consistent with my earlier results when the input characters were coded as intended by the authors of the source data, but which agreed with Farris when the reference point was chosen in the artificial manner he claims to have used. It is interesting that Farris insists on the one hand that he must use real data to obtain meaningful results, while on the other hand, he chooses to ignore that real data, instead choosing his reference point in an arbitrary manner. He also makes no mention of the extent to which his results produce useful classifications. Somewhat similar observations were made by Sokal and Rohlf at the NTI3 meeting that was held at Harvard October 26-7, 1979. I admit that my pair of examples proves nothing; they merely serve to indicate a possible explanation. The matter will be more extensively pursued in a later paper.

In closing I would like to mention one further reason why special similarity does not perform well. As I have done earlier, I am restricting my attention to binary characters in which 1 is the preferred state. The problem

might lie in the fact that special similarity not only ignores matched 0's, but it also ignores information provided by characters possessed by one but not both objects of a pair under consideration. Thus it treats equally the following 3 pairs of objects:

$$\begin{cases} A_1 & 1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0 \\ B_1 & 1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0 \end{cases}$$

$$\begin{cases} A_2 & 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 0 \\ B_2 & 1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0 \end{cases}$$

$$\begin{cases} A_3 & 1\ 1\ 1\ 1\ 1\ 1\ 0\ 0\ 0\ 0\ 0\ 0 \\ B_3 & 1\ 0\ 0\ 0\ 0\ 0\ 1\ 1\ 1\ 1\ 1\ 0 \end{cases}$$

But one can argue that (A_1, B_1) should be more similar than either of the other pairs. Similar examples can be constructed for characters having more than two states.

Acknowledgement. The programs needed for this paper were written by Zachary Smith, and the data processed on the University of Massachusetts Control Data Corporation CYBER 70 computer.

Note. For purposes of drawing the tree diagrams in the figures, each DC was rank ordered.

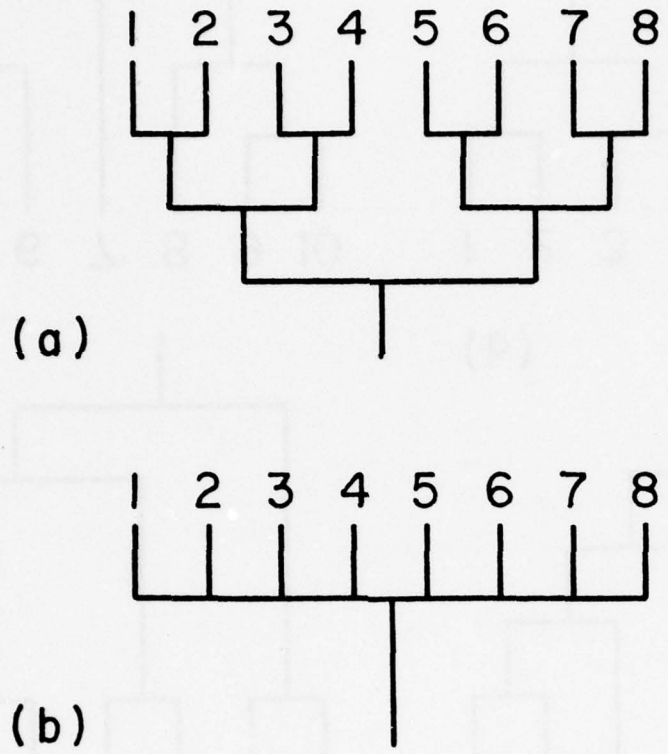


Figure 1. Two classifications of hypothetical data in table 1.

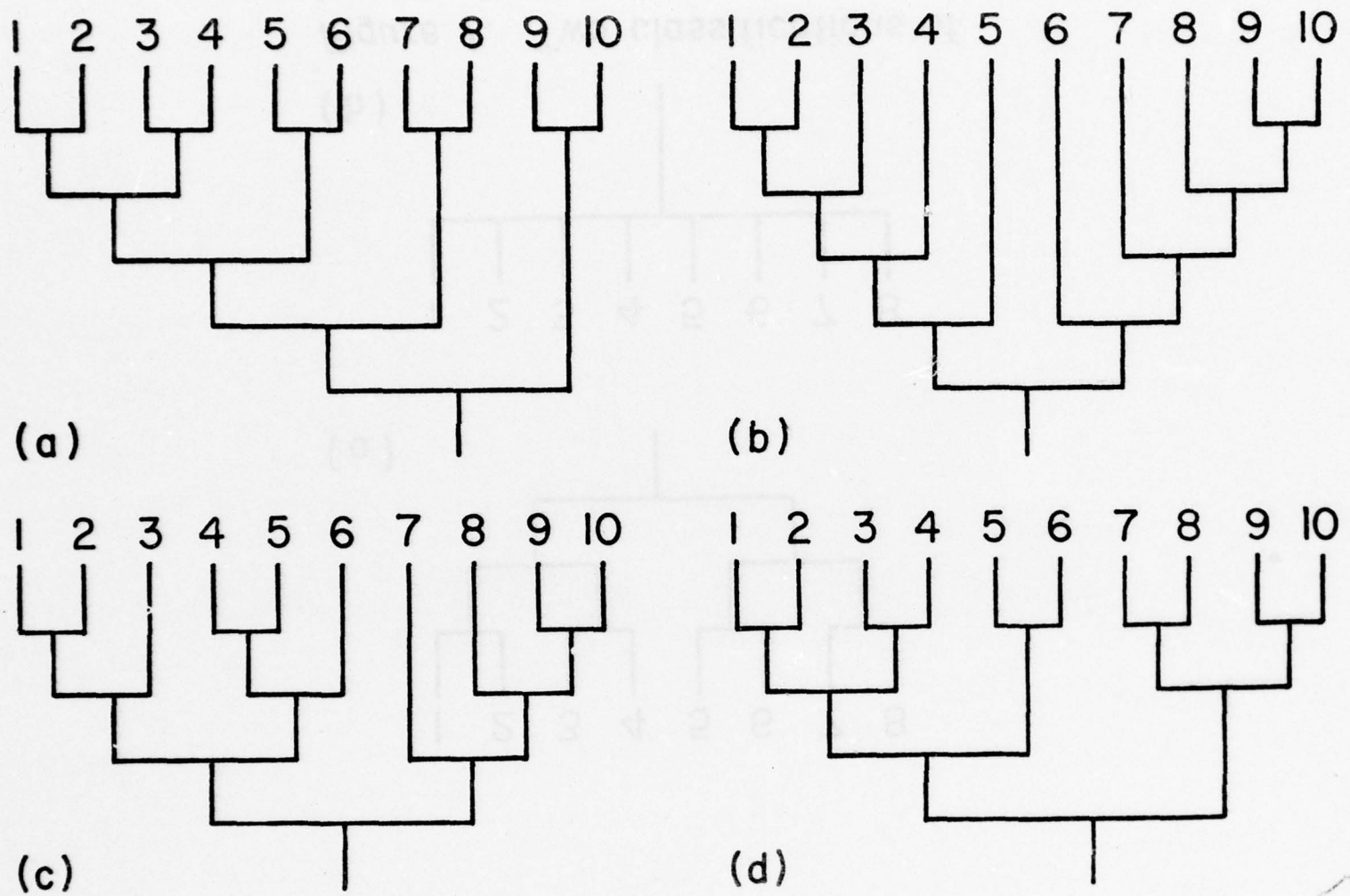


Figure 2. Test classifications.

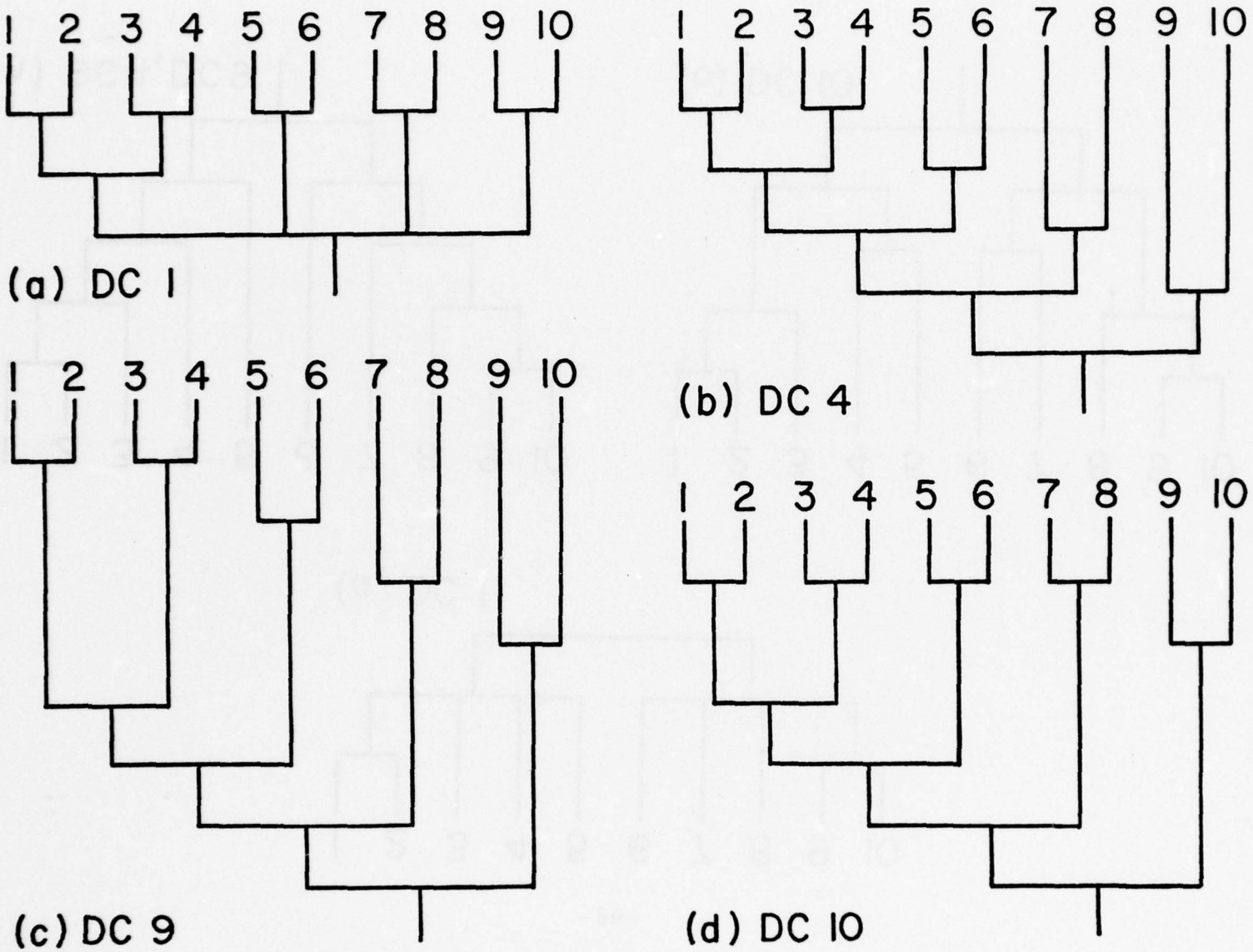


Figure 3. Classifications for data in table 6.

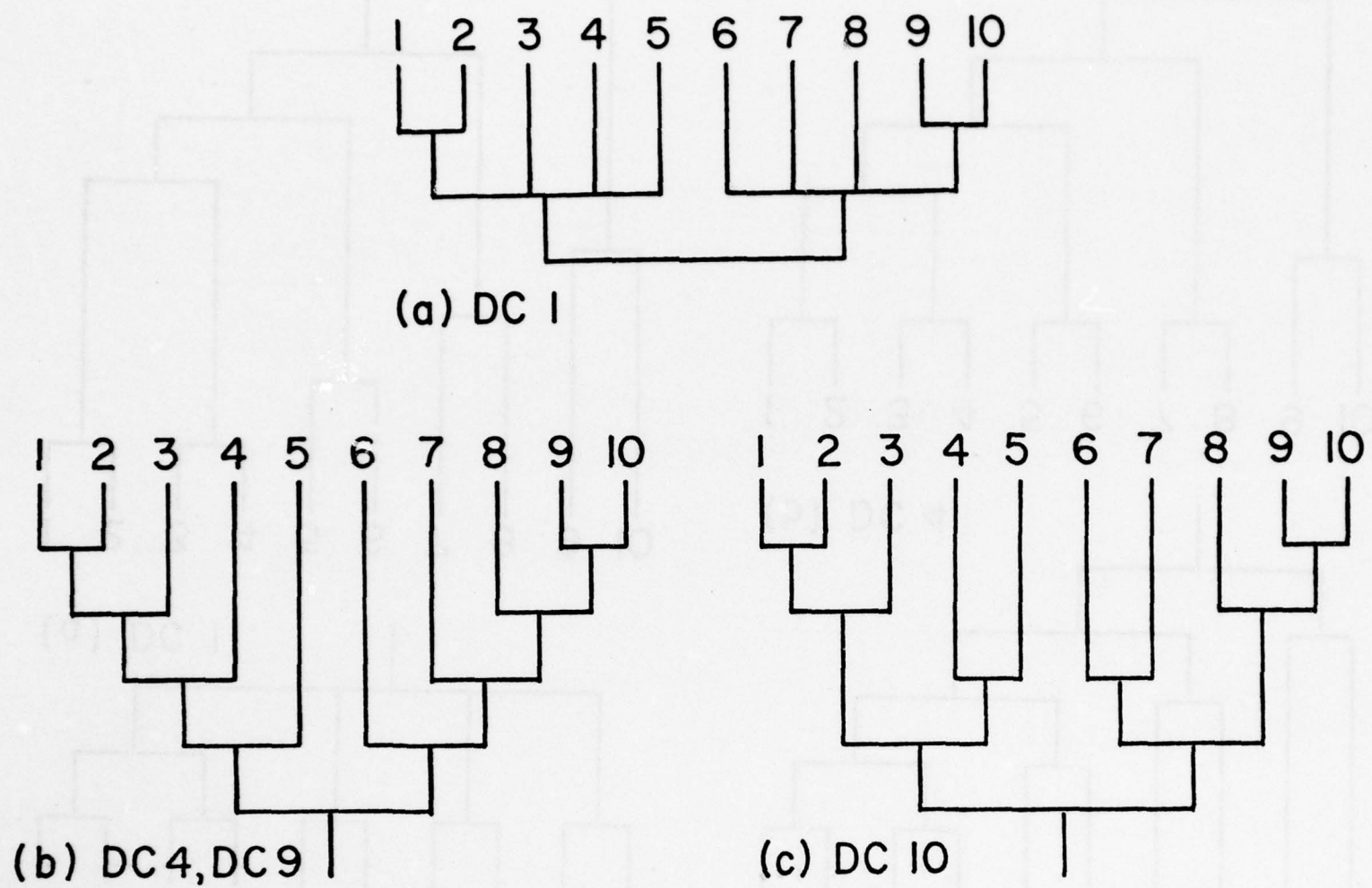


Figure 4. Classifications for data in table 7.

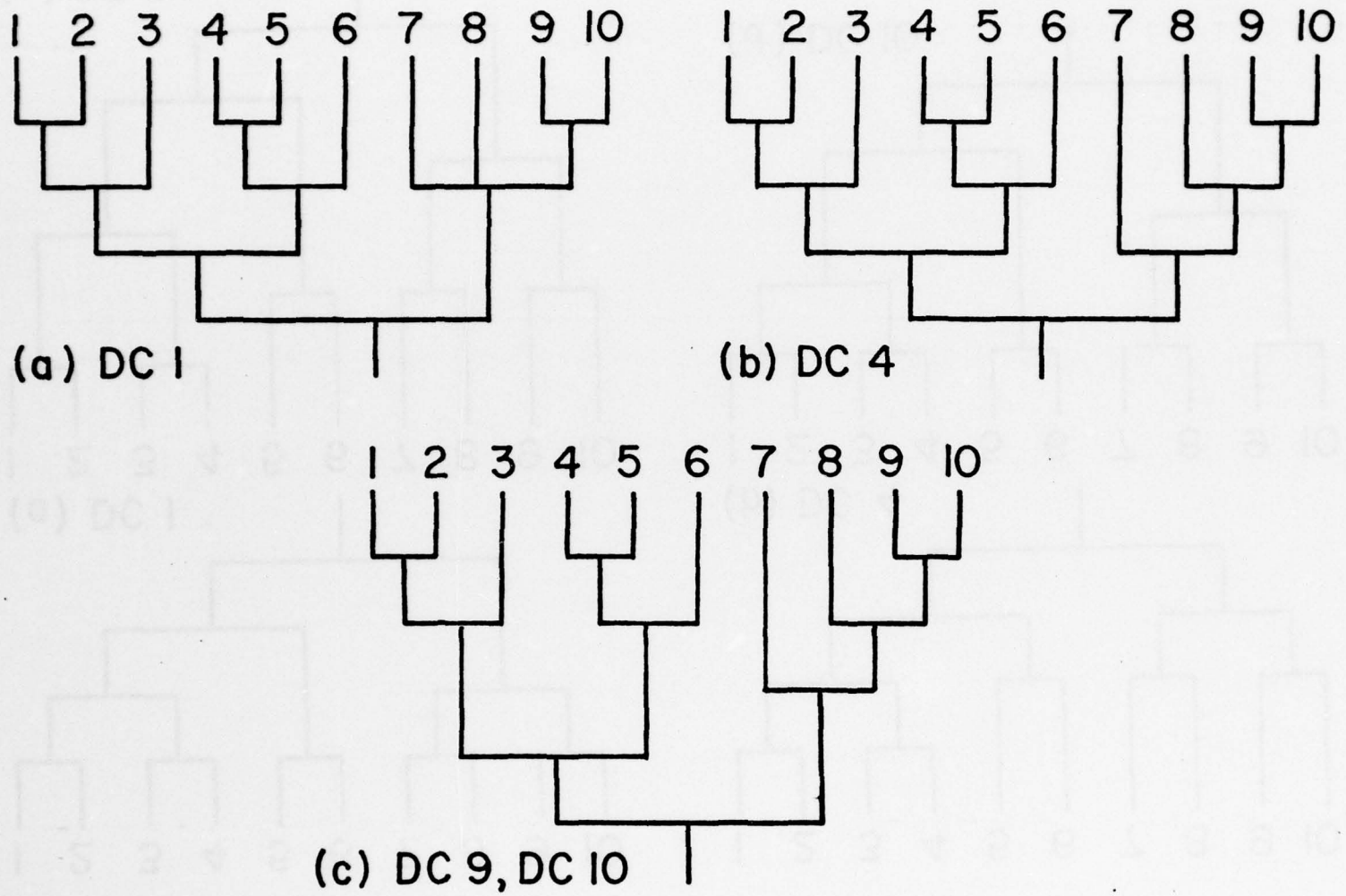


Figure 5. **Classifications for data in table 8.**

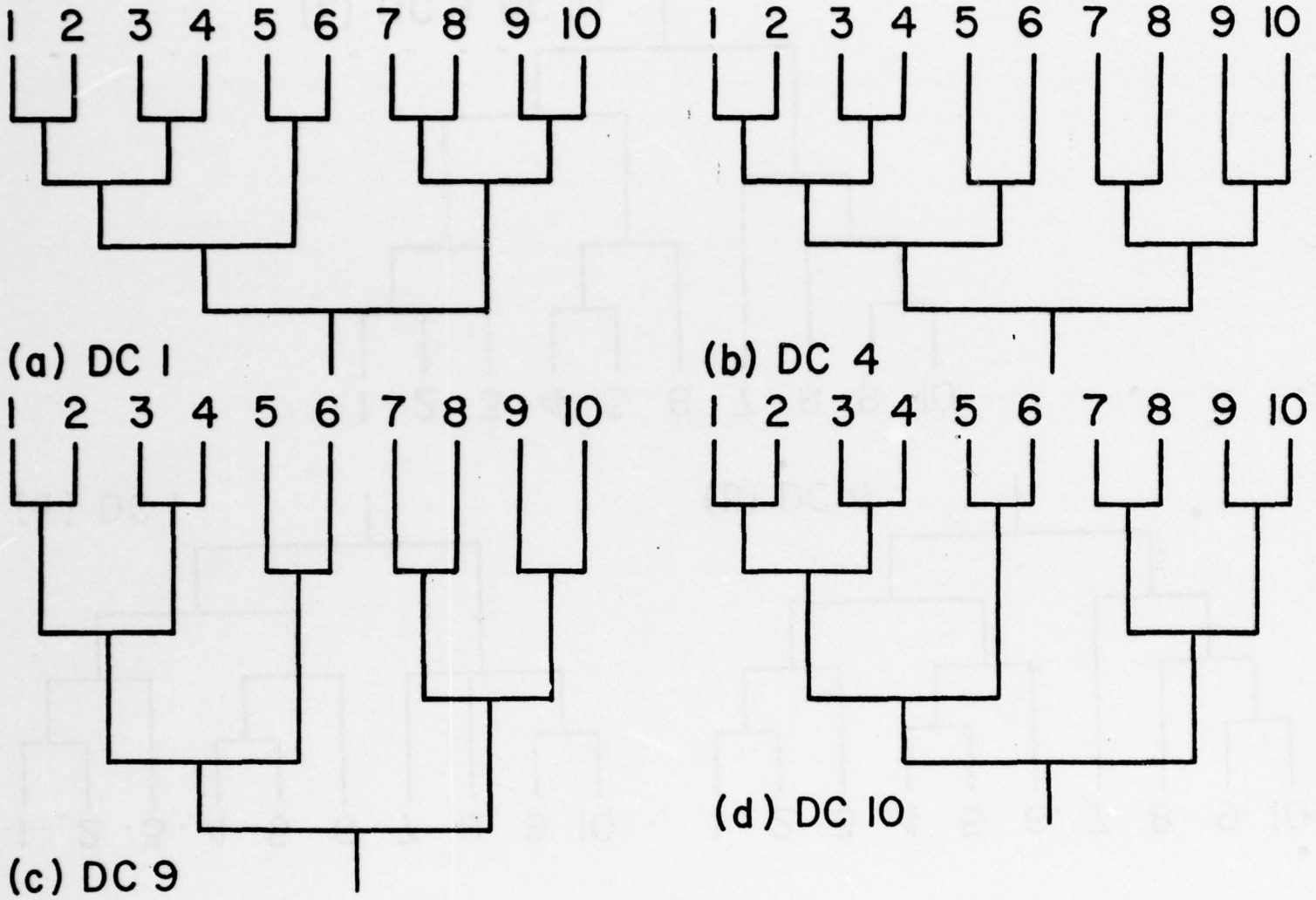


Figure 6. Classifications for data in table 9.

APPENDIX. LIST OF TABLES.

Table 1. Hypothetical Data Matrix.

From Farris (1979:201)

Character	Taxa								Factor
	1	2	3	4	5	6	7	8	
1	1	0	0	0	0	0	0	0	5
2	0	1	0	0	0	0	0	0	1
3	0	0	1	0	0	0	0	0	5
4	0	0	0	1	0	0	0	0	1
5	0	0	0	0	1	0	0	0	5
6	0	0	0	0	0	1	0	0	1
7	0	0	0	0	0	0	1	0	5
8	0	0	0	0	0	0	0	1	1
9	1	1	0	0	0	0	0	0	3
10	0	0	1	1	0	0	0	0	1
11	0	0	0	0	1	1	0	0	3
12	0	0	0	0	0	0	1	1	1
13	1	1	1	1	0	0	0	0	1
14	0	0	0	0	1	1	1	1	1

Table 2. List of Dissimilarity Measures

Measure	Formula	Name
DC1	$(b + c)/(a + b + c + d)$	Simple Matching Coefficient
DC2	$(b + c)/(a + b + c)$	Jaccard's Coefficient
DC3	$1 - a/(a + d + 2b + 2c)$	Rogers-Tanimoto
DC4	$1 - a/(a + b + c + d)$	Russel and Rao
DC5	$1 - 2a/(2a + 2d + b + c)$	
DC6	$1 - [a/(2a + 2c) + a/(2a + 2b)]$	Kulczynski
DC7	$1 - [a/(4a+4c) + a/(4a+4b) + d/(4b+4d) + d/(4c+4d)]$	
DC8	$1 - a^2/(a + b)(a + c)$	Ochia
DC9	$1 - (ad)^2/(a + b)(a + c)(d + b)(d + c)$	
DC10	$bc/(ad + bc)$	Yule
DC11	$.5 + (bc - ad)/2[(a+b)(a+c)(d+b)(d+c)]^{.5}$	Product-moment
DC12	$(b + c)/(a + b + c + \sqrt{ad})$	Baroni-Urbani and Nuser

Working with a fixed pair J,K of OTU's,

a = number of characters shared by J and K b = no. of characters possessed by J and not K

c = number of characters K has but not J d = no. of characters possessed by neither one

The measures were taken from Anderberg (1973) except for DC12 which appears in Baroni-Urbani and Nuser (1976).

Each measure was converted to a dissimilarity measure and then normalized so that it took values between 0 and 1.

Table 3. Summary of random data results. The figures relate to the percentage of success in 25 trials.

DC	T = 10 Mean	SD	T = 5 Mean	SD	T = 2 Mean	SD
1	.568	.4571	.6	.4163	.408	.3581
2	.6	.432	.688	.37	.592	.4222
3	.648	.3664	.712	.3516	.624	.4013
4	.608	.4778	.68	.3786	.6	.3958
5	.544	.3852	.656	.3343	.6	.4041
6	.616	.4394	.744	.3583	.576	.4136
7	.672	.4316	.712	.3516	.616	.3955
8	.600	.4339	.712	.3468	.608	.3936
9	.616	.42	.704	.3221	.592	.4327
10	.808	.3628	.856	.2973	.72	.4041
11	.656	.4224	.712	.3516	.608	.3936
12	.64	.4163	.68	.3916	.576	.3972

Table 4. Hypothetical Data Matrix

From Janowitz (1979:197)

	Characters						
Taxa	1	2	3	4	5	6	7
A	1	1	0	1	1	0	0
B	1	1	1	0	0	1	0
C	0	0	1	1	0	0	1

Table 5. Summary of results for data in Table 4. The figures relate to the percentage of success in 25 trials of 5 replications each.

DC	T = 10 Mean	SD	T = 5 Mean	SD	T = 2 Mean	SD
1	.04	.0817	.072	.1275	.2	.2
2	.016	.0554	.04	.0817	.112	.1641
3	.04	.0817	.056	.0917	.256	.1781
4	.088	.1013	.36	.2082	.36	.2082
5	.008	.04	.032	.0748	.112	.1641
6	0	0	.016	.05538	.112	.1641
7	.008	.04	.024	.0663	.12	.1528
8	.04	.08165	.032	.0748	.144	.1583
9	.24	.1732	.208	.1869	.208	.1869
10	.016	.0554	.008	.04	.096	.1541
11	.008	.04	.024	.0663	.12	.1528
12	.024	.0663	.04	.1	.04	.1

Table 6. Hypothetical Data Matrix Designed to Represent Classification Indicated in Fig. 2 (a)

OTU	Character																		
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1	1	0	0	0	0	1	1	1	1	1	0	0	0	0	0	0	0	0	0
2	1	0	0	0	0	1	1	1	1	0	1	0	0	0	0	0	0	0	0
3	0	1	0	0	0	1	1	1	1	0	0	1	0	0	0	0	0	0	0
4	0	1	0	0	0	1	1	1	1	0	0	0	1	0	0	0	0	0	0
5	0	0	1	0	0	0	1	1	1	0	0	0	0	1	0	0	0	0	0
6	0	0	1	0	0	0	1	1	1	0	0	0	0	0	1	0	0	0	0
7	0	0	0	1	0	0	0	1	1	0	0	0	0	0	0	1	0	0	0
8	0	0	0	1	0	0	0	1	1	0	0	0	0	0	0	0	1	0	0
9	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	1	0
10	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	1
Rep.	0	0	1	2	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 7. Hypothetical Data Matrix To Represent Classification in Fig. 2 (b)

	Character																		
OTU	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1	1	0	1	0	1	0	1	0	1	1	0	0	0	0	0	0	0	0	0
2	1	0	1	0	1	0	1	0	1	0	1	0	0	0	0	0	0	0	0
3	0	0	1	0	1	0	1	0	1	0	0	1	0	0	0	0	0	0	0
4	0	0	0	0	1	0	1	0	1	0	0	0	1	0	0	0	0	0	0
5	0	0	0	0	0	0	1	0	1	0	0	0	0	1	0	0	0	0	0
6	0	0	0	0	0	0	0	1	1	0	0	0	0	0	1	0	0	0	0
7	0	0	0	0	0	1	0	1	1	0	0	0	0	0	0	1	0	0	0
8	0	0	0	1	0	1	0	1	1	0	0	0	0	0	0	0	1	0	0
9	0	1	0	1	0	1	0	1	1	0	0	0	0	0	0	0	0	1	0
10	0	1	0	1	0	1	0	1	1	0	0	0	0	0	0	0	0	0	1
Rep.	0	0	0	0	0	0	0	0	0	0	0	1	2	3	3	2	1	0	0

Table 8. Hypothetical Data Matrix To Represent Classification in Fig. 2 (c)

	Character																		
OTU	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1	1	0	0	1	0	0	1	0	1	1	0	0	0	0	0	0	0	0	0
2	1	0	0	1	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0
3	0	0	0	1	0	0	1	0	1	0	0	1	0	0	0	0	0	0	0
4	0	1	0	0	1	0	1	0	1	0	0	0	1	0	0	0	0	0	0
5	0	1	0	0	1	0	1	0	1	0	0	0	0	1	0	0	0	0	0
6	0	0	0	0	1	0	1	0	1	0	0	0	0	0	1	0	0	0	0
7	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	1	0	0	0
8	0	0	0	0	0	1	0	1	1	0	0	0	0	0	0	0	1	0	0
9	0	0	1	0	0	1	0	1	1	0	0	0	0	0	0	0	0	1	0
10	0	0	1	0	0	1	0	1	1	0	0	0	0	0	0	0	0	0	1
Rep.	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	2	1	0	0

Table 9. Hypothetical Data Matrix To Represent Classification in Fig. 2 (d)

	Character																		
OTU	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1	1	0	0	0	0	1	0	1	1	1	0	0	0	0	0	0	0	0	0
2	1	0	0	0	0	1	0	1	1	0	1	0	0	0	0	0	0	0	0
3	0	1	0	0	0	1	0	1	1	0	0	1	0	0	0	0	0	0	0
4	0	1	0	0	0	1	0	1	1	0	0	0	1	0	0	0	0	0	0
5	0	0	1	0	0	0	0	1	1	0	0	0	0	1	0	0	0	0	0
6	0	0	1	0	0	0	0	1	1	0	0	0	0	0	1	0	0	0	0
7	0	0	0	1	0	0	1	0	1	0	0	0	0	0	0	1	0	0	0
8	0	0	0	1	0	0	1	0	1	0	0	0	0	0	0	0	1	0	0
9	0	0	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0	1	0
10	0	0	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	1
Rep.	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0

Table 10. Summary of results on addition of 3 random error characters

		Method 1		Method 2		Method 3		Method 4	
Table	DC	Mean	SD	Mean	SD	Mean	SD	Mean	SD
6	1	.9809	.01179	.8	.9189	.2	.6325	.0123	.0074
	4	.9237	.05477	2.2	1.033	.3	.6749	.1562	.0592
	10	.9818	.02123	.7	.8233	.1	.3162	.0117	.0058
7	1	.9789	.03481	.6	.6992	.1	.3162	.0296	.0098
	4	.9218	.03481	3.2	.7888	.6	.4889	.106	.0555
	10	.9869	.01639	.6	.6992	.2	.4216	.0296	.0109
8	1	.9503	.05002	1.3	1.419	.1	.3162	.0784	.1121
	4	.774	.1715	4.1	1.37	.7	.6749	.2389	.1403
	9	.9205	.0456	1.8	1.317	.6	.6992	.1327	.1435
	10	.9253	.1072	1.3	1.252	.3	.6749	.1078	.1326
9	1	.9738	.01322	1.4	1.075	.5	.527	.0317	.0159
	4	.8732	.1038	3.7	1.059	.8	.7888	.1568	.139
	9	.9397	.0342	1.3	1.16	.7	.6749	.0309	.0122
	10	.9899	.00769	.6	.6992	.4	.5164	.0259	.0116
1	1	.9166	.08901	1.7	1.703	.4	.6992	.0810	.0641
	4	.6235	.3698	3.7	2.058	.8	1.398	.3167	.2507
	9	.7797	.2466	1.8	2.044	.9	1.37	.1362	.1594
	10	.963	.03663	.8	1.135	.2	.4216	.0653	.0304

Note: See text for a description of the 4 methods. Observe that smaller values of Method 4 represent better results, while the opposite is true for the other methods.

Table 11. Number of times that DCi performed better than DCj on data from Table 10.

	Data From Table					Total	Total No. Trials
	6	7	8	9	1		
DC10 better than DC9			7	10	7	24	29
DC10 better than DC4	9	10	10	10	8	47	49
DC10 better than DC1	6	8	5	10	7	35	49
DC9 better than DC4			10	7	7	24	29
DC1 better than DC9			9	9	7	25	29
DC1 better than DC4	10	10	10	10	8	48	49

Note: One trial was omitted from the data of Table 1 because DC4 produced a constant output for which a product moment correlation could not be calculated. On this trial, the correlations for DC1, DC9 and DC10 were, respectively, .9217, .7889, and .9852.

Table 12. Summary of results on introduction of 5% random error

Data Table	DC	Method 1		Method 4	
		Mean	SD	Mean	SD
6	1	.9353	.04752	.1098	.1025
	4	.9745	.02951	.02838	.05097
	9	.9454	.04724	.02808	.05124
	10	.9667	.02846	.03788	.0569
7	1	.943	.02941	.07667	.07269
	4	.9404	.04181	.07576	.05765
	9	.9385	.0389	.06475	.06629
	10	.972	.0271	.06545	.06569
8	1	.9094	.07074	.1129	.09648
	4	.9142	.08961	.1048	.1148
	9	.9183	.03548	.08616	.08377
	10	.9466	.08528	.08758	.08279
9	1	.9539	.03724	.06768	.1067
	4	.9292	.07129	.09747	.101
	9	.9277	.04512	.06747	.1088
	10	.9608	.06278	.06768	.1067
1	1	.965	.03271	.0537	.06059
	4	.8864	.2406	.1085	.186
	9	.9477	.0427	.05212	.01389
	10	.986	.00922	.05265	.01483

Note: u = .6 clustering was used.

Table 13. Number of times that DCi performed better than DCj on data from Table 12.

	Data From Table					Total
	6	7	8	9	1	
DC10 better than DC9	9	10	9	9	8	45
DC10 better than DC4	3	7	9	7	8	34
DC10 better than DC1	10	9	8	8	8	43
DC9 better than DC1	7	4	5	2	4	22
DC4 better than DC9	10	7	7	6	6	36
DC4 better than DC1	9	6	6	3	5	29

Table 14. Summary of results on introduction of 15% random error (5 trials).

		Correlation	
Table	DC	Mean	SD
6	4	.6302	.1,35
	10	.7756	.2425
7	4	.6344	.2337
	10	.7155	.1532
8	4	.4095	.1554
	10	.4813	.3311
9	4	.6088	.178
	10	.6961	.2173
1	4	.2683	.2195
	10	.4181	.2957

Note: See Note for Table 12.

Table 15. Summary of results for simulation of composite Axiom A4 error (5 trials).

Data Table	DC	Method 1 Mean	SD	Method 4 Mean	SD
6	1	.8833	.1034	.1578	.1576
	4	.8385	.1769	.202	.1793
	9	.8611	.06172	.1246	.1072
	10	.865	.2045	.1402	.1552
7	1	.902	.05696	.1301	.1186
	4	.8506	.03606	.204	.08058
	9	.872	.04642	.1422	.1108
	10	.8612	.1312	.1416	.1136
8	1	.8445	.1705	.1727	.1995
	4	.8042	.1712	.2133	.2005
	9	.8546	.1209	.1562	.2083
	10	.8301	.2807	.1527	.2031
9	1	.9373	.0331	.06222	.03343
	4	.8841	.05317	.0998	.02252
	9	.8829	.05768	.06121	.03305
	10	.9174	.01604	.06303	.03531
1	1	.732	.1563	.2587	.1817
	4	.3895	.3085	.4619	.1965
	9	.6862	.1166	.1587	.1345
	10	.8542	.1197	.146	.1153

Table 16. Summary of results using cophenetic correlation coefficient,

Data Table	DC	3 error Mean	SD	5% error Mean	SD	Axiom A4 error Mean	SD
6	1	.9508	.0204	.9194	.0203	.8917	.05091
	4	.8798	.0698	.9424	.02204	.8486	.08884
	9			.9747	.01962	.9405	.03649
	10	.9155	.01998	.9078	.02902	.8394	.1163
7	1	.9416	.00673	.8876	.03746	.8578	.05188
	4	.8916	.02111	.9038	.02711	.8575	.05297
	9			.954	.01974	.9329	.02459
	10	.8873	.03547	.8688	.02597	.8158	.05825
8	1	.9	.03755	.8475	.05023	.8516	.06368
	4	.7803	.05838	.8713	.04383	.7906	.08261
	9	.933	.01441	.9402	.0207	.926	.03768
	10	.7994	.05249	.8433	.03849	.7788	.08113
9	1	.9267	.01839	.9198	.03162	.8919	.02071
	4	.8425	.05963	.8994	.04595	.8577	.04616
	9	.954	.01383	.9573	.0239	.9343	.003787
	10	.8769	.03566	.894	.0457	.8737	.03773
1	1	.8859	.06506	.8994	.0476	.7644	.06996
	4	.7749	.1085	.8426	.1722	.6129	.1331
	9	.9049	.08787	.9684	.02573	.8288	.09006
	10	.7936	.09229	.894	.0457	.7122	.1021

Table 17. Number of times that cophenetic correlation was higher for DCi than DCj on data with 3 error characters from Table 14.

	Data from Table					Total	Total Trials
	6	7	8	9	1		
DC9 better than DC1			10	10	7	27	29
DC9 better than DC4			10	10	8	28	29
DC9 better than DC10			10	10	8	28	29
DC1 better than DC4	10	10	10	9	6	45	49
DC1 better than DC10	10	10	10	9	7	46	49
DC10 better than DC4	9	4	7	7	5	32	49

Note: One trial was omitted from Table 1. See Table 11. On this trial the values for DC1, DC9 and DC10 were .7656, .8152 and .7663.

Table 18. Number of times that cophenetic correlation was higher for DCi than DCj on data with 5% error from Table 16.

	Data from Table					Total
	6	7	8	9	1	
DC9 better than DC1	10	10	10	10	10	50
DC9 better than DC4	10	10	10	10	10	50
DC9 better than DC10	10	10	10	10	10	50
DC1 better than DC10	7	9	5	8	5	34
DC4 better than DC10	10	10	7	4	5	36
DC4 better than DC1	9	6	7	2	3	27

Table 19. Correlation between results of Table 16 with results of Tables 10 and 12

Table	3 error char.	5% error data
6	.8093	.3148
7	.2502	-.07789
8	.6538	.07483
9	.459	.1509
1	.4175	.8228

Table 20. Cophentic Correlation for results of Table 14.

Table	DC	Correlation Mean	SD
6	4	.7156	.0533
	10	.7436	.05731
7	4	.7007	.1472
	10	.7472	.04824
8	4	.683	.06887
	10	.6576	.05599
9	4	.7257	.06824
	10	.7357	.0766
1	4	.5994	.1563
	10	.622	.07049

REFERENCES

Anderberg, M. R. 1973. Cluster analysis for applications. Academic Press, New York.

Baroni-Urbani, C. and M. W. Buser. 1976. Similarity of binary data. Syst. Zool. 25:251-259.

D'Andrade, R. G. 1978. U-statistic hierarchical clustering. Syst. Zool. 43:59-67.

Estabrook, G. F., C. S. Johnson and F. R. McMorris. 1976. A mathematical foundation for the analysis of cladistic character compatability. Math Biosci. 29:181-187.

Farris, J. S. 1977. On the phenetic approach to vertebrate classification. In Hecht, M. K., P. C. Goody and B. M. Hecht (eds.), Major patterns in vertebrate evolution. Plenum, New York, pp. 823-850.

Farris, J. S. 1979. On the naturalness of phylogenetic classifications. Syst. Zool. 28:200-214.

Ferris, S. D. and G. S. Whitt. 1978. Phylogeny of tetraploid catostomid fishes based on the loss of duplicate gene expression. Syst. Zool. 27:189-206.

Janowitz, M. F. 1979. A note on phenetic and phylogenetic classifications. Syst. Zool. 28:197-199.

Janowitz, M. F. 1979a. Monotone equivariant cluster methods. SIAM J. Appl. Math. 37:148-165.

Watson, L., W. T. Williams and G. N. Lance. 1966. Angiosperm taxonomy:a comparative study of some novel numerical techniques/ J. Linn. Soc. (Bot.) 59:491-501.

Department of Mathematics and Statistics
University of Massachusetts
Amherst, MA 01003